Exercise: Mobile phone data anonymization

Note
The purpose of this exericse is to illustrate how easy it can be to deanonymise data and how it can have real consequences. It is not meant as a guide for deanonymisation, but as a reminder to take good care of your data.

Data

In this exercise you will be working in a hypothetical situation where badly anonymised call detail record (CDR) data has either been leaked from or released by computational social science researchers. The data set in questions is a pseudonymised call data record set with the following characteristics:

> - The data consists of mobile phone calls collected by a mobile phone operator.
> - Each data entry records the caller, receiver, and time of the start of the call. Each line in the data file is one entry. (See example below.)
> - The data is pseudonymised by numbering the user identifications from 0 to N-1, where N is the number of users. We assume that this process is done properly in a way that there is no information related to the users in these pseudonyms: for example, the numbering is not based on the numerical order of the phone numbers of the users.
> - The data contains all calls between all the customers of the operator in a single city, with a total of around 100k customers.
> - The data consists of all calls between 2022-04-01 and 2022-05-11 within customers of the operator in the city, and in total there are around 1.4 million calls.

As an example, an entry in the data (i.e., a line in the data file) looks like:

```
61708 48191 2022-04-01 00:00
```

which indicates a phone call from user 61708 to user 48191 at 2022-04-01 00:00.

At first, it might seem that the pseudonymisation done here is enough to protect the identities in this data sets, but as we will soon see, it is definitely not!

The data set you will be working on in this exercise is synthetically created by the lecturer of this course (who has worked on similar but much larger data sets from a mobile phone operator). You can download the data set from here: cdr-2022-04-01-to-2022-05-11.txt . (In Chrome, directly clicking the link opens the text file in the browser. To download it, you can right-click the link and choose `Save Link As...`.)

The task

In this exercise, you will be playing the part of a malicious person taking advantage of the pseudonymised data set.

In this scenario, you are suspecting that Alice has been making regular secretive phone calls to someone, and want to know if this is the case and who this person is. You are suspecting that the calls might be to either Bob or Carol who you also know. You find an anonymized phone call data set that has been made publicly available in the internet, and decide to use that to investigate.

Your tasks:

a. Find youself in the data.

b. Find Alice, Bob, and Carol in the data.

c. Find who Alice has been calling with and when.

You start by looking at your phone and noting down these three phone calls in your call history:

- Call you made to Alice, 2022-05-01 at 09:14
- Call you made to Bob, 2022-04-16 at 10:20
- Call you made to Carol, 2022-04-21 at 11:43

Points 15 / 15 My submissions 1 / 3 Deadline Friday, 5 May 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 10 May 2023, 19:00).

Find yourself in the data

Question 1 15 / 15

Use the three time stamps of the calls you have made to find yourself in the data. What is your ID number in the data:

68819

Correct!

Submit

After you have found yourself in the data, you can easily identify everyone you have called in the data by comparing the timestamps in your call history to the ones you have made in the pseudonymised data.

Points 5 / 5 My submissions 1 / 3 Deadline Friday, 5 May 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 10 May 2023, 19:00).

Find Alice in the data

Question 1 5 / 5

What is the ID number of Alice in the data:

59715

Correct!

Submit

Points 5 / 5 My submissions 1 / 3 Deadline Friday, 5 May 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 10 May 2023, 19:00).

Find Bob in the data

Question 1 5 / 5

What is the ID number of Bob in the data:

60644

Correct!

Submit

Points 5 / 5 My submissions 1 / 3 Deadline Friday, 5 May 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 10 May 2023, 19:00).

Find Carol in the data

Question 1 5 / 5

What is the ID number of Carol in the data:

20924

Correct!

Submit

Now that you have identified you contacts in the data, you can start investigating the calls between them.

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 5 May 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 10 May 2023, 19:00).

Bob or Carol?

Question 1 10 / 10

Who has Alice been calling, Bob or Carol?

- (•) Bob
- ( ) Carol

Correct!

Submit

Points 5 / 5 My submissions 1 / 3 Deadline Friday, 5 May 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 10 May 2023, 19:00).

When was the first call?

Question 1 5 / 5

When was the first call that Alice made to Bob or Carol? Use the timestamp format from the CDR file: yyyy-mm-dd HH:MM

2022-05-01 03:45

Correct!

Submit

Points 5 / 5 My submissions 1 / 3 Deadline Friday, 5 May 2023, 19:00 To be submitted alone

The deadline for the assignment has passed (Wednesday, 10 May 2023, 19:00).

How many calls?

Question 1 5 / 5

How many calls were there in total between Alice and Bob or Carol?

6

Correct!

Submit

Further reading: https://www.pnas.org/content/113/20/5536.short